What is claimed is:

1. A system for providing capitalization correction for unstructured excerpts, comprising:

a preprocessor to tokenize an excerpt of unstructured content into a set of words; and

a capitalizer to analyze the set of words for correct capitalization, comprising:

an evaluator to evaluate individual characters constituting at least one such word in the set of words; and

a filter to skip the at least one such word if determined to be of a predefined type.

2. A system according to Claim 1, further comprising:

a title capitalizer to provide one or more of the words with an initial letter in uppercase and each remaining letter in lowercase.

3. A system according to Claim 1, further comprising:

a sentence capitalizer to provide only an initial such word with an initial letter in uppercase and each remaining letter in lowercase.

4. A system according to Claim 1, further comprising:

a word analyzer to skip at least one of each such word comprising a number, each such word including no vowels, and each such word not occurring at a start of a phrase and constituting at least one of an article, conjunction, preposition.

5. A system according to Claim 1, further comprising:

a lexicon comprising one or more reference words with at least one reference word defining a form of capitalization for the reference word;

a matcher to match the at least one such word against the reference words, the evaluator skipping each such word if a matching reference word is found.

6. A system according to Claim 1, further comprising:

a proper noun capitalizer to provide the individual letters in each such word comprising a noun with no vowels in uppercase.

7. A system according to Claim 1, further comprising:

a tokenizer to tokenize the excerpt into the one or more words and one or more punctuation marks.

8. A method for providing capitalization correction for unstructured excerpts, comprising:

tokenizing an excerpt of unstructured content into a set of words; and

analyzing the set of words for correct capitalization, comprising:

evaluating individual characters constituting at least one such word in the set of words; and

skipping the at least one such word if determined to be of a predefined type.

9. A method according to Claim 8, further comprising:

providing one or more of the words with an initial letter in uppercase and each remaining letter in lowercase.

10. A method according to Claim 8, further comprising:

providing only an initial such word with an initial letter in uppercase and each remaining letter in lowercase.

11. A method according to Claim 8, further comprising:

skipping at least one of each such word comprising a number, each such word including no vowels, and each such word not occurring at a start of a phrase and constituting at least one of an article, conjunction, preposition.

12. A method according to Claim 8, further comprising:

maintaining a lexicon comprising one or more reference words with at least one reference word defining a form of capitalization for the reference word;

matching the at least one such word against the reference words; and

skipping each such word if a matching reference word is found.

13. A method according to Claim 8, further comprising:

providing the individual letters in each such word comprising a noun with no vowels in uppercase.

14. A method according to Claim 8, further comprising:

tokenizing the excerpt into the one or more words and one or more punctuation marks.

15. A computer-readable storage medium holding code for performing the method according to Claim 8.

16. An apparatus for providing capitalization correction for unstructured excerpts, comprising:

means for tokenizing an excerpt of unstructured content into a set of words; and

means for analyzing the set of words for correct capitalization, comprising:

means for evaluating individual characters constituting at least one such word in the set of words; and

means for skipping the at least one such word if determined to be of a predefined type.

17. A system for building a lexicon for use in capitalization correction for unstructured excerpts, comprising:

a ripper assembling a list of word sets from unstructured content, each word set comprising a word and at least one variation on capitalization;

an aggregator aggregating each word set, comprising:

an analyzer identifying at least one word set comprising significant statistics; and

a non-standard capitalization selector selecting at least one such variation within the identified word set having a non-standard capitalization, and adding the at least one such variation to the lexicon.

18. A system according to Claim 17, further comprising:

a tokenizer tokenizing the excerpt into the one or more words and one or more punctuation marks.

19. A system according to Claim 18, wherein hyphenated words are split into a plurality of the words.

20. A system according to Claim 17, wherein at least one variation appearing at the start of a sentence is skipped.

21. A system according to Claim 20, wherein the non-standard capitalization comprises the at least one variation occurring in an excerpt having fewer than half of individual letters provided in uppercase.

22. A system according to Claim 17, further comprising:

a normalizer normalizing a plurality of the words extracted relative to a source of the structured Web content.

23. A system according to Claim 17, wherein the significant statistics comprises at least four occurrences of at least one such variation within a word set.

24. A system according to Claim 17, wherein the non-standard capitalization comprises the at least one variation having any individual letter other than the first individual letter provided in uppercase.

25. A system according to Claim 17, further comprising:

a standard capitalization selector selecting at least one such variation within the identified word set having a standard capitalization, and adding the at least one such variation to the lexicon.

26. A system according to Claim 17, further comprising:

a validator applying implicit rules for capitalization, and skipping each at least one variation subject to at least one such implicit rule.

27. A system according to Claim 26, wherein the implicit rules comprise skipping at least one variation based on position within a sentence or phrase.

28. A system according to Claim 26, wherein the implicit rules comprise at least one of a number, having no vowels, and constituting at least one of an article, conjunction and preposition.

29. A system according to Claim 26, wherein the implicit rules comprise normalizing a number of occurrences for each at least one variation using at least one of a normalizing function and relative to a source of the at least one variation.

30. A system according to Claim 26, wherein the implicit rules comprise accommodating multiple forms of capitalization for each at least one variation by annotating each capitalization form with a frequency count and skipping those of the each at least one variation occurring infrequently.

31. A system according to Claim 17, further comprising:

a hash table maintaining the lexicon.

32. A system according to Claim 31, further comprising:

at least one record specifying at least one such word as a key into the hash table, and associating at least one such variation within the word set as a preferred capitalization.

33. A method for building a lexicon for use in capitalization correction for unstructured excerpts, comprising:

assembling a list of word sets from unstructured content, each word set comprising a word and at least one variation on capitalization;

aggregating each word set, comprising:

identifying at least one word set comprising significant statistics;

selecting at least one such variation within the identified word set having a non-standard capitalization; and

adding the at least one such variation to the lexicon.

34. A method according to Claim 33, further comprising:

tokenizing the excerpt into the one or more words and one or more punctuation marks.

35. A method according to Claim 34, further comprising:

splitting hyphenated words into a plurality of the words.

36. A method according to Claim 33, further comprising:

skipping at least one variation which may be at the start of a sentence.

37. A method according to Claim 36, wherein the non-standard capitalization comprises the at least one variation occurring in an excerpt having fewer than half of individual letters provided in uppercase.

38. A method according to Claim 33, further comprising:

normalizing a plurality of the words extracted relative to a source of the structured Web content.

39. A method according to Claim 33, wherein the significant statistics comprises at least four occurrences of at least one such variation within a word set.

40. A method according to Claim 33, wherein the non-standard capitalization comprises the at least one variation having any individual letter other than the first individual letter provided in uppercase.

41. A method according to Claim 33, further comprising:

selecting at least one such variation within the identified word set having a standard capitalization, and adding the at least one such variation to the lexicon.

42. A method according to Claim 33, further comprising:

applying implicit rules for capitalization; and

skipping each at least one variation subject to at least one such implicit rule.

43. A method according to Claim 42, wherein the implicit rules comprise skipping at least one variation based on position within a sentence or phrase.

44. A method according to Claim 42, wherein the implicit rules comprise at least one of a number, having no vowels, and constituting at least one of an article, conjunction and preposition.

45. A method according to Claim 42, wherein the implicit rules comprise normalizing a number of occurrences for each at least one variation using at least one of a normalizing function and relative to a source of the at least one variation.

46. A method according to Claim 42, wherein the implicit rules comprise accommodating multiple forms of capitalization for each at least one variation by annotating each capitalization form with a frequency count and skipping those of the each at least one variation occurring infrequently.

47. A method according to Claim 33, further comprising:

maintaining the lexicon structured as a hash table.

48. A method according to Claim 47, further comprising:

specifying at least one such word as a key into the hash table; and

associating at least one such variation within the word set as a preferred capitalization.

49. A computer-readable storage medium holding code for performing the method according to Claim 33.

50. An apparatus for building a lexicon for use in capitalization correction for unstructured excerpts, comprising:

means for assembling a list of word sets from unstructured content, each word set comprising a word and at least one variation on capitalization;

means for aggregating each word set, comprising:

means for identifying each word set comprising significant statistics;

means for selecting at least one such variation within the identified word set having a non-standard capitalization; and

means for adding the at least one such variation to the lexicon.